Human Detection with a Multi-sensors Stereovision System

Yannick Benezeth, Pierre-Marc Jodoin, Bruno Emile, Hélène Laurent, Christophe Rosenberger

Human detection with a multi-sensors stereovision system

Y. Benezeth[1], P.M. Jodoin[2], B. Emile[3], H. Laurent[4], and C. Rosenberger[5]

[1] Orange Labs, 4 rue du Clos Courtel, 35510 Cesson-Sévigné - France
[2] MOIVRE, Université de Sherbrooke, BP 91226, Sherbrooke, J1K 2R1 - Canada
[3] Institut PRISME, Université Orléans, IUT de l'indre, 36000 Châteauroux - France
[4] ENSI de Bourges, Institut PRISME, 88 bd Lahitolle, 18020 Bourges Cedex - France
[5] GREYC, ENSICAEN, Université de Caen - CNRS, 14000 Caen - France
yannick.benezeth@orange-ftgroup.com

Abstract. In this paper, we propose a human detection process using Far-Infrared (FIR) and daylight cameras mounted on a stereovision setup. Although daylight or FIR cameras have long been used to detect pedestrians, they nonetheless suffer from known limitations. In this paper, we present how both can collaborate inside a stereovision setup to reduce the false positive rate inherent to their individual use. Our detection method is based on two distinctive steps. First, human positions are detected in both FIR and daylight images using a cascade of boosted classifiers. Then, both results are fused based on the geometric information of the sterovision system. In this paper, we present how human positions are localized in images, and how the decisions taken by each camera are fused together. In order to gauge performances, a quantitative evaluation based on an annotated dataset is presented.

1 Introduction

Techniques for locating humans in still images and videos have long been studied. It is now used in all kinds of military and civilian applications such as surveillance and security applications, energy-saving control on air-conditioning and lighting in offices, or simply counting people entering and leaving a building, to name of few [1]. But detecting humans in real-life scenarios is a fundamentally hard problem to solve, mainly because of the wide variability of human postures. Furthermore, small object size, occlusion, bad weather conditions, sudden illumination changes, camouflage problems, and the need for real-time applications, are common challenges one has to deal with.

As mentioned by Gavrila in its review paper [1], human detection methods can be divided in three categories namely (1) the 3D approaches, (2) the 2D approaches with explicit shape model, and (3) the 2D approaches without explicit shape model. Methods in (1) and (2) look forward at recovering 3D/2D body parts and posture on pre-localized blobs [2] often resulting in a stick-figure representation. While these methods do not require a training phase, they have strong assumptions on the content of the scene and often require the use of non-trivial

mathematical models. On the other hand, methods in (3) detect people without explicitly locating their body parts. In fact, based on a training database, these methods extract features (*e.g.* edges, gradients, shape, wavelet coefficients, etc.) and, following a clustering step (*e.g.* SVM, Adaboost, etc.) separate human from non-human shapes [3], [4].

Most of the methods presented above were meant to work on daylight cameras (or visible). However, as their cost keep decreasing, far-infrared (FIR) cameras (often called IR or thermic cameras) gain more interest for human detection [5], [6], [7] as they provide numerous advantages (night vision, relatively uniform backgrounds, etc.). However, as opposed to daylight cameras, FIR cameras fail at detecting humans in hot summer days and often suffer from floor reflection. Note that a study on human detection in FIR and daylight images can be found in [8].

As false detections in FIR and daylight images are due to different causes, it can be interesting to make them collaborate in a stereovision framework so that the number of false detections inherent to their individual use is reduced. To our knowledge, very few papers have addressed that issue before, the closest one being by Bertozzi *et al.* [9]. However, their method uses two pairs of stereovision systems from which they match disparity maps. In our method, only two cameras are required and the detection is based on a machine learning method and a fusion step. Each video stream of the stereovision system is independently processed based on Viola *et al.*'s method [4] before their results are merged together. The paper is organized as follows: we first present the used stereovision system and the implemented human detection system. Performances and possible applications of the system are finally presented.

2 THE STEREOVISION SYSTEM

The objective of our method is to combine information from a FIR and a daylight camera mounted side by side. But prior to do so, lets first review some basic stereovision notions. It is well known that, given the epipolar geometry of a stereovision system, each point observed in one image corresponds to an epipolar line in the other image. Such correspondance is typically determined via the so-called fundamental matrix [15]. The fundamental matrix F is a 3×3 matrix satisfying the relation $x_A'^T F x_A = 0$ in which A is a point in the world reference frame imaged as x_A in the first view and x'_A in the second. A simple projection $l_A = F x_A$ permits to determine the corresponding epipolar line l_A of x_A in the second image. This point-to-line relation stipulates that points lying on the epipolar line l_A in camera 2 are the only ones that can match x_A in camera 1.

Many methods have been proposed to estimate the fundamental matrix between two cameras, the most commonly implemented being the 8-points algorithm [15]. This being said, the fundamental matrix can also be obtained by combining the extrinsic and calibration matrices of each camera:

$$F = [P'C]_x P' P^+ \tag{1}$$

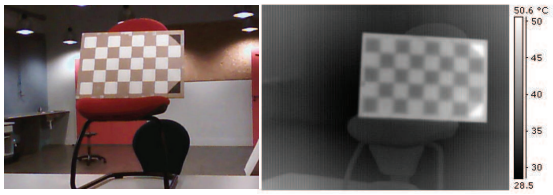

Fig. 1. Camera calibration setup: images acquired by the daylight and FIR cameras.

where P and P' are the projection matrices of the first and second camera, P^+ is the pseudo-inverse of P and C is the camera center. By their very nature, P and P' are obtained by multiplying together the calibration and extrinsic matrices (see chap 9 of [15] for more details). The reason for using the calibration and extrinsic matrices in our setup is two fold: estimate F following Eq. 1 to enforce the human detection procedure and 2) estimate the 3D position of each detected person (localization examples will be presented in section 4).

The intrinsic and extrinsic parameters are always estimated following a calibration procedure involving a pattern with known dimensions [11]. In our method, the picture of a calibration pattern is taken simultaneously by both cameras. Note that by its very nature, the FIR camera needs an atypical pattern with "warm" and "cold" areas (see Fig.1). This is achieved with a heat lamp mounted atop the calibration pattern so its dark squares are heated up.

3 HUMAN DETECTION METHODS

As mentioned previously, images from the FIR and daylight cameras are first independently processed. Then, each detection (bounding box) in one camera is matched (or confirmed) with the detections in the other camera. In this section, we describe how the human detection method works on still images and how results from both cameras are fused together.

3.1 Detection on FIR and Daylight images

In order to detect humans in FIR and daylight images, we use Viola *et al.*'s method [4]. Here, 14 Haar-like filters are used and, as shown in Fig 2, those filters are made of two or three black and white rectangles. The feature values x_i are computed with a weighted sum of pixels of each component.

Each feature x_i is then fed to a simple one-threshold weak classifier f_i :

$$f_i = \begin{cases} +1 & \text{if } x_i \geq \tau_i \\ -1 & \text{if } x_i < \tau_i \end{cases} \tag{2}$$

where $+1$ corresponds to a human shape and -1 to a non-human shape. The threshold τ_i corresponds to the optimal threshold that minimizes the misclassification error of the weak classifier f_i estimated during the training stage. Then, a

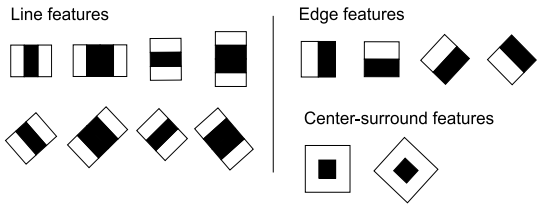

Fig. 2. Haar-like filters used by our human detection method.

more robust classifier is built with several weak classifiers trained with a boosting method [12]:

$$F_j = sign(c_1 f_1 + c_2 f_2 + \ldots + c_n f_n). \tag{3}$$

Then, a cascade of boosted classifiers is built. F_j correspond to the boosted classifier of the j^{th} stage of the cascade. Each stage can reject or accept the input window. Whenever an input window passes through every stages, the algorithm labels it as a human shape. Note that humans are detected in a sliding window framework [4].

3.2 Detection with the stereovision system

As mentioned in section 2, knowing the epipolar geometry of the sterovision system, one can link a point in one image to its corresponding epipolar line in the second image. In our method, this correspondence is used to confirm every human shape detected in one camera with those detected in the other camera.

Lets consider that M human shapes have been detected in the daylight image and N have been detected in the FIR image. As shown in Fig.3, lets also consider that A_i^{vis} and B_i^{vis} are the top-left and bottom-right points of the i^{th} human shape in the daylight image (represented by a bounding box) and dA_i^{vis} and dB_i^{vis}, their respective epipolar lines in the FIR image obtained with the fundamental matrix. In our method, a detected shape $i \in [1, M]$ in the daylight image is kept if and only if there is a shape $j \in [1, N]$ such that the distance between A_j^{FIR} and dA_i^{vis} and between B_j^{FIR} and dB_i^{vis} is smaller than a predefined threshold (obtained empirically). Whenever that test fails, the detected shape is deleted. In Fig. 3, two human shapes have been detected in the daylight image and only one in the FIR image. In this example, only shape 1 has been kept.

Of course, this algorithm is used both ways such that the shapes in the daylight image are confirmed with those in the FIR image and vice versa.

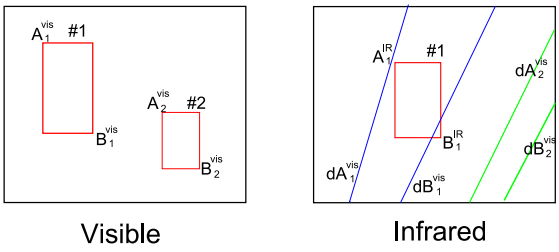

Fig. 3. Example of decision fusion based on epipolar geometry.

4 EXPERIMENTAL RESULTS

4.1 Datasets

Since a classifier is used in each spectrum, two learning datasets are required (cf. examples in Fig.4). The positive datasets are made of 1208 daylight images (from [3]) and 1175 FIR images (from the OTCBVS dataset [13, 14] and our manually annotated images). Both negative learning datasets are made of 3415 gray-level images (FIR and Daylight).

Fig. 4. Example of FIR and daylight images from the learning dataset.

The human detection rate with FIR and daylight cameras have been evaluated with several real-life videos taken in three different areas. That test dataset is made of 640 daylight and FIR images randomly extracted from videos and manually annotated.

4.2 Results

In order to gauge performance, Precision/Recall curves are used :

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP}, \qquad \text{Recall} = \frac{\#TP}{\#TP + \#FN}, \tag{4}$$

where $\#TP$, $\#FP$ and $\#FN$ stands for the number of true positives, false positives and false negatives respectively. We present in Fig. 5 results obtained

using the fundamental matrix to perform the point-to-line correspondance. The *FIR* and *Daylight* curves show results obtained after processing the daylight and FIR images outside our stereovision setup. As one might expect, results are far more precise with the FIR images. This is because FIR images have more uniform backgrounds and their human shape *vs.* background contrast is much stronger than in daylight images. The *Improved daylight* and *Improved FIR* curves show results obtained with the fusion process described in 3.2. As can be seen, our method clearly improves human detection performances in both spectrums.

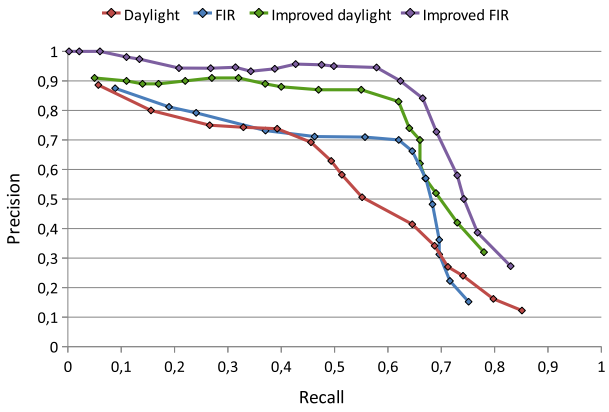

Fig. 5. Precision/Recall curves for Daylight and FIR images processed with and without our stereovision setup.

Detection examples are presented in Fig. 6. On the first row, one can see detections in a FIR image and a daylight image with false positives. On the second row, detections have been fused with our method and false positives (the reflection on the floor on the left and the chair on the right) have been deleted. Note that our fusion procedure is very fast since it only requires two projections per detection.

As explained in section 2, by using the extrinsec and intrinsec parameters of the calibrated scameras, it is possible to locate humans in 3D and so to verify their activities. We present in Fig. 7 two localization results. In the first example, the rectangle represents the ground truth location and the crosses correspond to the estimated location. The person coordinates are obtained retrieving the coordinates of the detection centroid in both spectrum and then calculating its 3D coordinates using stereo-triangulation. In the second example, we show the estimated path of a person drawn into the floor map. In this figure, the individual enters from the right door, stays for a while near the shelf, and then leaves through the same door. We can note that this information is sufficient for home care applications where a more precise location would not be useful.

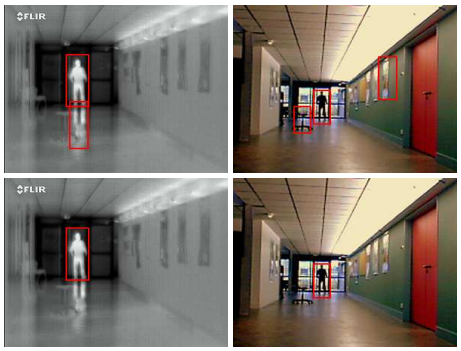

Fig. 6. First row: detection examples. Second row: detection examples with the stereovision system.

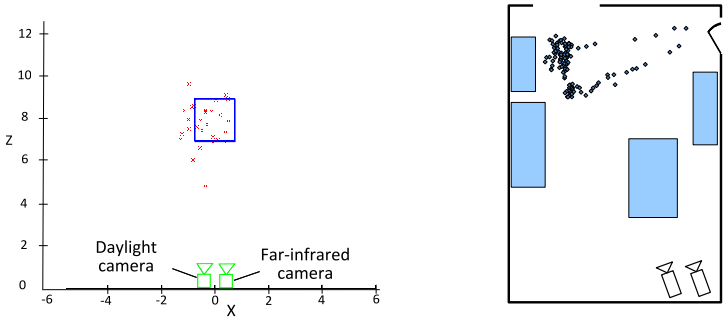

Fig. 7. Localisation results (the X and Z axis are in meters).

5 CONCLUSION

In this paper, we present a stereovision-based human detection system which uses a far-infrared and daylight camera. Since the FIR and a daylight cameras suffer from different limitations, our method combines both cameras in order to reduce the number of false positives inherent to their individual use. In our method, human positions are first detected in each camera independently with a cascade of boosted classifiers. Then, the detected shapes of each spectrum are fused in order to remove false positives. Results have been quantitatively shown with an evaluation study based on various manually annotated real-life videos. We have shown that our system distinctively outperforms the classic human detection.

The aim of this work is to decrease the number of false detections taking into account information given by our stereovision system. In the future, we plan to study benefits of decreasing the number of missed detections taking into account temporal information, for example integrating this human detection stereovision system in a simple tracking framework. Then, we will consider a fusion scheme

combining detection results of both cameras in the world reference frame instead of doing fusion in the images reference frames. So we will obtain one single detection result at each time for the whole system.

References

1. D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey", *in Proc. of CVIU*, vol. 73, pp. 82–98, 1999.
2. C.R. Wren, A. Azarbayejani, T. Darrell and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *in IEEE Trans. PAMI*, vol. 19, pp. 780–785, 1997.
3. N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance", *in Proc. of ECCV*, vol. 2, pp. 428–441, 2006.
4. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *in IEEE Proc. CVPR*, pp. 511–518, 2001.
5. M. Bertozzi, A. Broggi, A. Lasagni and M. Del Rose, "Infrared Stereo Vision-based Pedestrian Detection", *in Proc. of IVS*, pp. 24–29, 2005.
6. F. Xu, X. Liu and K. Fujimura, "Pedestrian Detection and Tracking With Night Vision", *in IEEE Trans. ITS*, vol. 6, pp. 63–71, 2005.
7. Y. Benezeth, B. Emile, H. Laurent and C. Rosenberger. "A Real Time Human Detection System Based on Far-Infrared Vision", *in ICISP*, pp 76–84, 2008.
8. Y. Fang, K. Yamada, Y. Ninomiya, B. Horn and I. Masaki, "Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection", *in Proc. of IVS*, pp. 505–510, 2003.
9. M. Bertozzi, A. Broggi, M. Felisa, G. Vezzoni and M. Del Rose, "Low-level Pedestrian Detection by means of Visible and Far Infra-red Tetra-vision", *in Proc. of IVS*, pp. 231–236, 2006.
10. A. Bovik, "Hand Book of Image and Video Processing", Academic Press, 2000.
11. J.Y. Bouguet and P Perona, "Closed-form Camera Calibration in Dual-space Geometry", *in Proc. of the ECCV*, 1998.
12. R.E. Schapire, "The boosting approach to machine learning: An overview," *in Workshop on N.E.C.*, 2002.
13. J. Davis and M. Keck, "A two-stage template to person detection in thermal imagery", *in Workshop on A.C.V*, 2005.
14. J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery", *in CVIU*, vol. 106 , pp. 162–182 2007.
15. R. Hartley and A. Zisserman, "Multiple view geometry in computer vision", Cambridge University Press, 2000.